DOCUMENT RESUME

ED 332 677 IR 014 994

AUTHOR Baker, Eva L.; Butler, Frances A.
TITLE Artificial Intelligence Measurement System, Overview and Lessons Learned. Final Project Report.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
SPONS AGENCY Advanced Research Projects Agency (DOD), Washington, D.C.
PUB DATE Feb 91
CONTRACT N00014-86-K-0395
NOTE 31p.
PUB TYPE Information Analyses (070) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Artificial Intelligence; *Cognitive Processes; *Comparative Testing; Evaluation Methods; Expert Systems; Language Processing; *Performance; Visual Perception

ABSTRACT

This report summarizes the work conducted for the Artificial Intelligence Measurement System (AIMS) Project which was undertaken as an exploration of methodology to consider how the effects of artificial intelligence systems could be compared to human performance. The research covered four areas of inquiry: (1) natural language processing and understanding; (2) expert systems; (3) machine vision and visual perception; and (4) technology assessment and evaluation. The four areas are discussed in turn with information provided regarding the goals of individual research efforts within each area. Comparative tests between human and computer performances are noted. A list of 31 project reports and 12 technical reports is included in the document. Names of project staff and consultants and a distribution list are appended. (DB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Final Project Report

ED332672

ARTIFICIAL INTELLIGENCE MEASUREMENT SYSTEM

OVERVIEW AND LESSONS LEARNED

Eva L. Baker

Frances A. Butler

February 1991

Artificial Intelligence Measurement System
Contract Number N00014-86-K-0395

Principal Investigator: Eva L. Baker

Center for Technology Assessment
UCLA Center for the Study of Evaluation

This research report was supported by contract number N00014-86-K-0395 from the Defense Advanced Research Projects Agency (DARPA), administered by the Office of Naval Research (ONR), to the UCLA Center for the Study of Evaluation. However, the opinions expressed do not necessarily reflect the positions of DARPA or ONR, and no official endorsement by either organization should be inferred. Reproduction in whole or part is permitted for any purpose of the United States Government.

Approved for public release; distribution unlimited.

IR014994

BEST COPY AVAILABLE

Artificial Intelligence Measurement System

Overview and Lessons Learned*

The Artificial Intelligence Measurement Systems (AIMS) project was undertaken as an exploration of methodology to explore how the effects of artificial intelligence systems could be compared to human performance. It was designed under a number of assumptions. First, that human performance is infinitely richer than the relatively primitive systems so far designed. Although the principal measurement strategy proposed treating system performance as if were a point in a distribution of human performance, there was no intention of equating conceptually computer systems and individual human performance. Prior research by Clancey (1988) for example, documented the fact that computer systems because of their consistency and dependence upon a coherent view (an expert) could be compared to a set of humans working on problems in a particular domain. Rather the exploratory goal of this project was to investigate whether intelligent systems could be placed on a continuum of human performance. In practice, this mapping would test some a priori correspondences, in that relatively unsophisticated systems would be mapped on a sample of individuals with relatively low performance and more sophisticated systems would map to individuals with more sophisticated levels of performance. If such a set of rough correspondences could be established, then it would be theoretically possible to benchmark systems under development in terms of progressively higher performing populations of individuals. Effectiveness, in terms of a performance and investment ratio, could be judged for increasingly expensive implementations. As a simple example, we could imagine comparing the mathematics problems solved by a system with the performance of students in kindergarten, 6th grade, and beginning calculus. Originally, the project was formulated to focus in one area--natural language understanding with the corresponding human performance domain of reading comprehension. This area held much promise because of (1) the rich research in both natural

*Citation not included in the references are in the list of project reports immediately following the reference page.

language understanding and reading comprehension and (2) the clear differentiation of individuals in terms of dimensions underlying text understanding. However, we were encouraged to consider multiple areas simultaneously, natural language understanding, including interfaces and texts, expert system shells and expert systems, and machine vision. The project also included a technology assessment component to permit reflection on our processes in the light of progress made elsewhere.

Another assumption of this project was that it would in part depend upon collaboration with members of the computer science discipline. It was also assumed that this requirement would provide a challenge because the form of evaluation we were exploring would not be within the expectations or values of members of this discipline. Although we experienced difficulties in acquiring systems for use and in sustaining interest of some computer scientists, critical components of this work were led or strongly influenced by members of the computer science community. Moreover, the project had a desired effect in energizing members of the community to explore approaches beyond standard software metrics to evaluate the impact of their efforts. The project experienced all the usual difficulties in dealing with complex software--delays in hardware implementations, concerns about the proprietary nature of code, as well as some unanticipated problems, such as the requirement but inability to evaluate systems implemented in classified domains. Staff also needed to quell occasional anxiety attacks related to imagined litigation occasioned by the public evaluation of commercial products.

As a strategy, the project invested the bulk of its resources in the natural language area. There it focused on two different types of implementations: interfaces that served to query databases or as front ends to expert systems and experimental text understanding systems. A principal effort in this project component was the development of a compatible descriptive/empirical strategy. The creation of a sourcebook of problems in natural language (Read, Dyer, Baker, Mutch, Butler, Quilici, & Reeves, 1990) was undertaken as a way to describe and map the field. This system could provide an interpretative context for the understanding of any empirical benchmarking results. Thus, the empirical benchmarking of systems could be

understood in terms of the difficulty of the task. A partial analogy is the degree of difficulty score paired with the performance score for a diver. Description was also a key element in the other project components as well, although no where was the effort as extensive as in the natural language understanding tasks. The machine vision project also created a sourcebook of problems (Skrzypek, Mesrobian, & Gungner, April 1988) and described existing vision systems and measures (Skrzypek, Mesrobian, & Gungner, March 1988). The expert system project created a framework for both expert systems and analogous human processes.

The empirical, human benchmarking strategy was predicated on the idea that existing tests would be available for administration, and that these existing, commercially available or research validated achievement tests would allow the benchmarking (or comparison) of multiple implementations. Early on in the project, it became clear that except in the area of vision, existing tests would be largely inappropriate because they did not reflect the domain specificity of particular implementations. Although linking and equating strategies are available to combine information from disparate tests, they imposed constraints in terms of the underlying dimension to be measured as well as required large sample sizes. Some existing measures were used, for example, standardized measures of reading ability, to assess performance differences, but for the most part, an unanticipated effort needed to be made in test development to create the performance base for comparison. This development proceeded according to strategies identified in Hively, Patterson, and Page (1968) and in Baker and Herman (1983) using what is known as domain referenced achievement tests. In the natural language area, an attempt was made to overcome the domain specifity problem. We created a measure that dissociated the structure of the query from its content base. This seemed to be the only approach available since we were assessing a system that needed to be reimplemented in each particular content domain each time it was applied, and the domain under development involved a classified Navy domain of information. In other test development, we were able to sidestep the domain issue by focusing on process, for example, the development of a test of metacognitive strategy described in the expert system component.

However, for much of our effort we were very much focused on the domain of task, the particular texts in systems or the particular content area of an expert system.

The project explored whether human benchmarking of computer systems is possible in a variety of classes of systems. Our answer to that question is yes. A corollary question is whether benchmarking processes are routinely feasible as evaluation procedures for intelligent systems. At the present time, our answer is no, for the practical and technical reasons above. We recommend the creation of descriptive resources, such as the Sourcebook, to enable the field to inform itself and keep abreast of the progress made by the community. Such resources could break down the unintentional barriers created by lineages of training or location. We further recommend the pursuit of benchmarking when there are sufficient implementations in a common area to support the investment in their common evaluation. Such evaluation would identify the differential emphases and effects of such systems in terms of their stated goals and in terms that program managers and policymakers could understand, that is, in terms of what ordinary or extraordinary people can and cannot do on their own.

Natural Language Understanding

Our research in the area of natural language understanding focused on methods of evaluating natural language processing (NLP) systems. Our goal in this area was two-fold:

1) we were interested in the identification and classification by example of problems in natural language understanding, and

2) we were interested in the development of an evaluation methodology which considers system output relative to or benchmarked to human performance.

The first approach took into account the processes that lead to output; the second approach was concerned with output only. These two evaluation metrics can be used to describe NLP systems in complementary ways. Baker (1987), Read, Dyer, and Feifer (1988), and Hecht and Wittrock

(1988) provide preliminary overviews of the issues addressed in the individual studies in the natural language understanding portion of the project.

Identification of Problems in Natural Language Understanding

The first approach to the issue of NLP system evaluation, that of identification by example and classification of problems in natural language understanding, is realized in practical form in the Natural Language Sourcebook (Read, Dyer, Baker, Mutch, Butler, Quilici, & Reeves, 1990). The Natural Language Sourcebook is a collection of 197 examples of natural language processing problems organized by a classification scheme which reflects an artificial intelligence perspective and cross-referenced by two other classification schemes, one reflecting a linguistic perspective and the other a cognitive-psychological perspective on the types of issues presented in the examples.

The Sourcebook developmental process involved a search through the artificial intelligence, computational linguistics, and cognitive science literature to identify examples of processing problems. Each example served as the basis for a Sourcebook entry. The entries, called "exemplars," each consist of 1) one or more sentences, a fragment of dialogue, or a piece of text which illustrates a conceptual issue, 2) a reference, and 3) a discussion of the problem a system might have in understanding the example. An example is used to illustrate each problem, but it is the discussion that defines the type of problem by delineating the information-processing issues involved. The Sourcebook exemplars provide discussions of concrete processing problems in terms of the general principles at issue. This grounding of the general in the specific makes the Sourcebook a uniquely useful and appropriate tool for evaluation of NLP systems.

At two different stages, the Sourcebook underwent rigorous content review. First, when 50 exemplars had been compiled, the Sourcebook was reviewed internally at UCLA by a linguist and a cognitive scientist. Then when 150 exemplars had been developed, the Sourcebook was sent for external review to experts in artificial intelligence and computer science at Carnegie Melon University, the University of Michigan, and the Illinois Institute of Technology. Based on

reviewer comments at both stages, substantive revisions were made in the Sourcebook, and additional exemplars were developed. Once the exemplars were completed, the linguistic and cognitive-psychological cross-indexing was added.

Finally, an electronic version of the Sourcebook database was developed (Herl, August 1990 and September 1990). This electronic HyperCard version of the Natural Language Sourcebook capitalizes on the modular structures of the Sourcebook exemplars and facilitates use of the multiple classification schemes by links between specific cards (exemplars). The HyperCard version of the Natural Language Sourcebook is accompanied by a user's manual (Herl, August, 1990).

The Sourcebook project is covered in Dyer and Read (1988) as well as in the introduction to the Sourcebook itself (Read et al., 1990). The cognitive-psychological classification scheme used for cross-referencing the Sourcebook exemplars is presented in Wittrock (1989). A status report on the Sourcebook was presented at the ONR contractor's meeting held at Princeton University, March 1990 (Butler & Baker, 1990).

An initial test of the usefulness of the Natural Language Sourcebook as a tool for describing and evaluating NLP systems is described in Mutch, 1990. This report provides an empirical verification of the problem coverage in the Natural Language Sourcebook by referencing output from one intelligent computer system, IRUS, to the Sourcebook exemplars. From the consideration of the IRUS queries in relation to the Natural Language Sourcebook, it appears that the coverage of processing problems presented in the Sourcebook is sufficiently comprehensive to be of practical use.

Benchmarking to Human Performance

The second approach to the issue of NLP system evaluation, that of evaluating NLP systems by benchmarking to human performance, was explored in two major studies. The first provides an initial specification of a continuum of difficulty for language a syntactic shell interface, IRUS, can process (Baker, Turner, & Butler, 1990). The continuum of difficulty is based on the

performance of kindergartners and first graders on comprehension tasks syntactically parallel to those accomplished by IRUS. Baker and Lindheim (1988) and Baker, Lindheim and Skrzypek (1988) provide preliminary descriptions of the study presented in Baker et al. (1990).

The second study provides a comparison of the abilities of six text understanding systems to answer specific questions about given texts with the abilities of humans to answer the same questions about the same texts (Butler, Baker, Falk, Herl, Jang, & Mutch, 1990). In this study, systems were benchmarked to grade equivalent groups of human subjects.

In Baker et al. (1990), correct responses for the human subjects were determined by how IRUS responded to parallel items (i.e., all the IRUS responses were taken to be correct), whereas in Butler et al. (1990), correct responses for both human subjects and intelligent computer systems were determined by the consensus responses of adult native speakers.

Baker et al. (1990) provides an initial verification of the feasibility of distinguishing intelligent computer system responses to natural language processing tasks by human developmental criteria; Butler et al. (1990) extends this initial investigation by looking at a larger range of human developmental stages and by actual benchmarking of systems' overall and differential capabilities to human capabilities as they vary with development.

Expert System Shells

This component of the project attempted to investigate reasonable approaches to the evaluation of expert system shells. It attempted to explore:

1) what methodologies available from social science might be brought to bear on the study of expert system shells;

2) what was the feasibility of implementing these strategies in a routine way because of commercial interests in shell quality.

This project began with the analysis of costs and benefits of experimental approaches to the study of expert systems, particularly the construction of an experiment manipulating shells and

tasks and assigning them to system developers with various levels of expertise. Even if critical variables, such as order, domain knowledge, and task generalizability could be controlled, the approach was rejected because of feasibility concerns--time, cost, and the small likelihood that system developers appropriate to represent the population of interest could be released from their regular tasks in order to complete our experimental requirements.

Instead, we decided to take a different tack and assess qualitatively the process of knowledge engineering and system development using a case study approach. Following a review of the literature (reported in Novak, Baker, & Slawson, 1991), the project recognized that typical software metrics in use for shell evaluation did not focus on in detail the processes nor the outcomes of development. Although our literature review did turn up studies focused on user satisfaction, and consumer guide sorts of analyses, in depth studies of knowledge engineering processes had not been made. Consequently, the project posited the idea of developing a 2x2 design for the conduct of intensive case studies, with one factor focusing on the sophistication of the shell in terms of representation and inferencing strategies and the other factor focusing on the nature of the problem, whether it was well defined or ill-structured. To undertake this work, a well defined problem, selecting the appropriate reliability index for use with a particular form of achievement test, was formulated. An expert psychometrician was identified and video tapes and observations of the knowledge engineering process were made. The first system employed was relatively unsophisticated, M-1™. The knowledge engineer had some previous domain knowledge and had experience in implementing other expert systems in this shell. The knowledge engineer prepared reports (Li, 1987; Li, 1988) and early progress in this effort was reported by Slawson, Novak, and Hambleton (1988). The implementation was reviewed by the expert and found to be unsatisfactory because of domain misconceptions by the knowledge engineer. Rather than proceed to completion, the expert recommended that we try something else. Principally using the existing videotapes and with minimal visits with the expert, another implementation of an expert system was made using NEXPERT™. At that point, given the difficulty and cost of this

strategy, with the approval of our advisors, we decided to focus on expert systems. The summary report of effort in this area is provided in Novak, Baker and Slawson, 1990.

Benchmarking Expert Systems

The problem of human benchmarking in an expert system context was addressed by research attending to the following questions:

1) What descriptive analyses of computer expert processes and human cognitive processes should be attempted?
2) On what dimensions could expert system performance be benchmarked on humans?

This work was conducted in cooperation with a subcontract to the Cognitive Science Laboratory of USC. The project initiated with a literature review of benchmarking of expert systems (O'Neil, Ni & Jacoby, 1990) in which it became clear that the project could opt to have computer-science driven models or psychologically driven models of benchmarking. Although it would be ideal to cross validate these approaches, we were constrained by the lack of availability of expert system implementations which would permit multiple tests of a psychological driven measurement model. The decision was to conduct human benchmarking according to the conceptual model originally outlined in Baker (1987), that is to norm an expert system's performance on samples of individuals. Expert systems always involve considerable amounts of domain-specific knowledge, thus, unlike the IRUS work described above, it was difficult to isolate the structure of tasks from content. We believed however we could, through the use of metaphor, transform the essence of an expert system (GATES, a system that assigned airplanes to gates in major airline hubs) into a valid psychological construct. The GATES program schedules by assigning an item to time, location, etc, without violating constraints. The psychological equivalent of this task is called self-monitoring in the literature. We surveyed extant measurement literature to identify an existing, high quality instrument to assess this aspect of human metacognition. When we found no such instrument, one was developed. Thus a study was designed that incorporated

both the benchmarking of outcomes (how well samples of students completed the GATES tasks) and of human processes (how well students planned, selected strategies, and monitored their behavior while conducting the task, and how aware they were of their processes). The design methodology both in the general case and as it applied to GATES is included in the report by O'Neil, Ni, Jacoby, and Swigger (1990). Finally, a report of the evaluation, using both process and outcome measures was prepared, following the conduct of experimental trials (O'Neil, Baker, Jacoby, Ni, & Wittrock, 1990). The methodology was demonstrated to be successful in that individuals with a priori different ability levels performed predictably. A summary of the entire set of activities is provided by O'Neil (1990).

Additional outcomes for this component of the project were found. One spin-off study looked at the applicability of current research in software engineering, human performance measurement, simulation, and machine learning for the evaluation of expert systems and suggested incorporating some of the techniques into a formal assessment methodology. The methodology was then applied to the GATES system (Swigger, O'Neil, Ni, & Jacoby, 1990). A second spin-off study investigated the GATES task as it provided an environment for the experimental test of explanation facilities. In an experiment, goals, tasks, and explanation types were manipulated (Jacoby, 1990). Probably the most important outcome was the development of apparently highly reliable and valid measures of human metacognition. These measures were developed using tested models from the realm of personality measurement, that is, both the trait of metacognition and its application under particular states were measured. Trait measurement means how an individual normally functions whereas state measures ask for his/her retrospective report of function under specific conditions. These measures are currently being experimentally employed in other performance assessment contexts (Baker & O'Neil, 1991). They seem to have promise as measures of engagement and attention to complex tasks, measures with obvious application to military and civilian training and to educational outcome assessment in general.

Machine Vision

The machine vision benchmarking component was completed under the direction of Dr. Josef Skrzypek of the UCLA Computer Science Department. This component sought to answer the following questions:

1. As a long term goal, the project investigated how machine vision might proceed as a joint effort between the neurosciences and computer science.

2. Specifically related to this project, the component sought to generate a framework for evaluating progress in machine vision by documenting the status of the field and investigating the human visual performances that could be benchmarked on a vision system?

The strategy used for the vision benchmarking component, initially described in Baker (1987) and Baker, Lindheim, and Skrzypek (1988) in some ways paralleled the strategy used in the natural language component. Three reports provide initial exploration of the machine vision strategy (Mesrobian & Skrzypek, June 1987; Paik, Gungner, & Skrzypek, June 1987; and Skrzypek & Mesrobian, November 1987). Following a conference of experts in computer science, neuroscience, and psychology, the project conducted an extensive reviews of 15 vision systems in order to identify possible categories along which machine vision systems could be evaluated. In the report by Skrzypek, Mesrobian, and Gungner (March 1988), each of these analyses is followed by justifications for the use of the human visual system as a model for a general purpose vision system. The report identifies visual tasks from existing tests and discusses them in terms of their corresponding computational neural substrates. Comparisons among systems are made along five dimensions: 1) image attributes; 2) perceptual primitives; 3) knowledge base; 4) object representation; and 5) control. Skrzypek and his colleagues rejected the attempt to benchmark individual vision systems directly. They did so for a number of reasons. One constraint was the idiosyncratic platforms used in the development of such systems. The cost of acquiring such sufficient hardware appropriately configured was well beyond the resources of this project. Similarly, the particular domain of interest for these systems was extremely narrow. When approaching the problem from the human side, benchmarking ran into some limitations, in

large measure because the bulk of existing systems focused on lower and middle range visual tasks with minimal cognitive demands. Such tasks, were outside accessible ranges for typical individuals. Simple tasks were automatic, e.g., matching to samples used in manufacturing systems, that people had no awareness of when and how they completed such tasks and one would need to drop to visually impaired or individuals with specific brain dysfunctions, caused by age, accident, or disease. On the other end, computer image enhancement pushed beyond the limits of individual capability. Instead, the team decided to work in the opposite direction. They created a model of general purpose vision. They assembled typical visual tasks provided to individuals in regular psychological tests, such as paper folding and block tests, and documented neuroscience evidence connected to them. Finally, they created a Sourcebook (Skrzypek, Mesrobian, & Gungner, April 1988) documenting data level visual tasks. Each entry consists of a problem statement, a discussion, references from the literature and examples.

Technology Assessment

A final component of this effort was the attempt to be reflective and self-conscious about the strategies we undertook to evaluate complex systems. These strategies involve technical, social, financial and policy dimensions. One integrative analysis of the problem where this project is used as an example was created by Baker (in press) from an invited chapter presented at a symposium on intelligent systems sponsored by the Air Force Human Resources Laboratory. As a culmination to the project, a conference was held at UCLA inviting a wide range of individuals from the military, academic and industrial sectors (Baker, Butler, & O'Neil, 1990). Each presentation was focused on either general models for assessing technology, cumulative findings in an area, and particular examples. Papers written by external consultants are included in the report. Because we are attempting to secure a commercial contract for the publication of these and redrafts of project reports, we prefer to restrict their circulation at this time (Baker, Butler, & O'Neil, 1991). The conference proved to be very much work-in-progress in its focus and

underscored the relatively little systematic thought given to the assessment (and evaluation) of technologies of all sorts. Clearly, working on the boundaries among fields, computer science, military training, education, evaluation, and psychometrics will provide a continuing challenge.

Summary

The AIMS project provided documentation of explorations of the benchmarking of intelligent systems on human performance. The project used both descriptive and empirical strategies and a wide range of methodologies. The project was conducted in the following areas: natural language understanding, expert systems, machine vision, and included a technology assessment component.

References

Clancey, W. J. (1988). Acquiring, representing, and evaluating a competence model of diagnostic strategy. In M.T.H. Chi, R. Glaser, & M. J. Farr (Eds.), The nature of expertise (pp. 343-418). Hillsdale, NJ: Erlbaum.

Hively, W., Patterson, H. L., & Page, S. A. (1968). A "universe-defined" system of achievement tests. Journal of Educational Measurement, 5, 275-290.

Baker, E. L., & Herman, J. (1983). Task structure design: Beyond linkage. Journal of Educational Measurement, 20 (2), 149-164.

Baker, E. L. (in press). Technology assessment: Policy and methodological issues for training. In H. Burns, C. Luckhardt, & J. Parlett (Eds.), Knowledge architectures in intelligent tutoring systems. Hillsdale, NJ: Erlbaum.

Baker, E. L., & O'Neil, H. F., Jr. Plan for NAEP motivation studies. Center for the Study of Evaluation, University of California, Los Angeles and University of Southern California.

Artificial Intelligence Measurement System

Project Reports

Contract No. N00014-86-K-0395

Principal Investigator: Eva L. Baker

Center for Technology Assessment
UCLA Center for the Study of Evaluation

February 1991

Natural Language Understanding

1. Baker, E. L. March 1987. Artificial Intelligence Measurement System (Briefing Charts). ONR Contractors' Meeting, Yale University.

2. Baker, E. L., & Lindheim, E. L. May 1988. A Contrast Between Computer and Human Language Understanding. CSE Technical Report 287. Center for Technology Assessment, UCLA Center for the Study of Evaluation.

3. Baker, E. L., Lindheim, E. L., & Skrzypek, J. May 1988. Directly Comparing Computer and Human Performance in Language Understanding and Visual Reasoning. CSE Technical Report 288. Center for Technology Assessment, Graduate School of Education, and Artificial Intelligence Laboratory, Computer Science Department, UCLA.

4. Baker, E. L., Turner, J. L., & Butler, F. A. March 1990. An Initial Inquiry into the Use of Human Performance to Evaluate Artificial Intelligence Systems. Center for Technology Assessment, UCLA Center for the Study of Evaluation.

5. Butler, F. A., & Baker, E. L. March 1990. Natural Language Sourcebook Status Report (Briefing Charts). ONR Contractors' Meeting, Princeton University.

6. Butler, F. A., Baker, E. L., Falk, T., Herl, H., Jang, Y., & Mutch, P. September 1990. Benchmarking Text Understanding Systems to Human Performance: An Exploration. Center for Technology Assessment, UCLA Center for the Study of Evaluation.

7. Dyer, M., & Read, W. April 1988. A Sourcebook Approach to Evaluating Artificial Intelligence Systems. Paper presented at the annual meeting of the American Educational Research Association. New Orleans.

8. Hecht, B. F., & Wittrock, M. April 1988. Cognitive and Linguistic Perspectives on Natural Language Understanding. Paper presented at the annual meeting of the American Educational Research Association. New Orleans.

9. Herl, H. August 1990. User's Manual, HyperCard Database for the Natural Language Sourcebook . Center for Technology Assessment, UCLA Center for the Study of Evaluation.

10. Herl, H. September 1990. Designing a HyperCard Database for the Natural Language Sourcebook. Center for Technology Assessment, UCLA Center for the Study of Evaluation.

11. Mutch, P. August 1990. Processing Problems in the IRUS Queries: An Empirical Verification of Problem Coverage in the Natural Language Sourcebook. Center for Technology Assessment, UCLA Center for the Study of Evaluation.

12. Read, W., Dyer, M., & Feifer, R. April 1988. What's So Hard About Understanding Language? Paper presented at the annual meeting of the American Educational Research Association. New Orleans.

13. Read, W., Dyer, M., Baker, E., Mutch, P., Butler, F., Quilici, A., & Reeves, J. 1990. Natural Language Sourcebook. Center for Technology Assessment, Center for the Study of Evaluation and Computer Science Department, UCLA.

14. Wittrock, M. C. June 1989. A Classification of Sentences Used in Natural Language Processing in the Military Services. CSE Technical Report 294. Center for Technology Assessment, UCLA Center for the Study of Evaluation.

Expert Systems

15. Jacoby, A. October 1990. Expert System Explanation: The User Perspective. Center for Technology Assessment, UCLA Center for the Study of Evaluation.

16. Li, Z. October 1987. An Expert System for Selecting the Index of Reliability. Department of Instructional Psychology and Technology, School of Education, University of Southern California.

17. Li, Z. April 1988. Knowledge Engineering Report: An Expert System for Selecting Reliability Index. Department of Instructional Psychology and Technology, School of Education, University of Southern California.

18. Novak, J. R., Baker, E. L., & Slawson, D. A. January 1991. The Evaluation of Expert System Shells. Center for the Technology Assessment, UCLA Center for the Study of Evaluation.

19. O'Neil, H. F., Jr., Ni, Y., & Jacoby, A. January 1990. Literature Review: Human Benchmarking of Expert Systems. Cognitive Science Laboratory, University of Southern California and Center for Technology Assessment, UCLA Center for the Study of Evaluation.

20. O'Neil, H. F., Jr., Ni, Y., Jacoby, A., & Swigger, K. M. September, 1990. Human Benchmarking Methodology for Expert Systems. Cognitive Science Laboratory, University of Southern California; Center for Technology Assessment, UCLA Center for the Study of Evaluation, and Department of Computer Sciences, University of North Texas.

21. O'Neil, H. F., Jr., Baker, E.L., Jacoby, A., Ni, Y., & Wittrock, M. October 1990. Human Benchmarking Studies of Expert Systems. Cognitive Science Laboratory, University of Southern California and Center for Technology Assessment, UCLA Center for the Study of Evaluation.

22. Slawson, D. A., Novak, J., & Hambleton, R. K. April 1988. A Qualitative Approach to the Evaluation of Expert System Shells. Paper presented at the annual meeting of the American Educational Research Association. New Orleans.

23. Swigger, K. M., O'Neil, H. F., Jr., Ni, Y., & Jacoby, A. October 1990. Assessment of Expert Systems. Department of Computer Sciences, University of North Texas; Cognitive Science Laboratory, University of Southern California, and Center for Technology Assessment, UCLA Center for the Study of Evaluation.

24. O'Neil, H. F., Jr. November 1990. Measurement of Expert Systems Effectiveness, Final Report. Cognitive Science Laboratory, University of Southern California.

Machine Vision

25. Mesrobian, E., & Skrzypek, J. June 1987. Discrimination of Natural Textures: A Neural Network Architecture. Paper presented at the Institute of Electrical and Electronics Engineers Annual International Conference on Neural Networks, San Diego.

26. Paik, E., Gungner, D., & Skrzypek, J. June 1987. UCLA SFINX--A Neural Network Simulation Environment. Paper presented at the Institute of Electrical and Electronics Engineers Annual International Conference on Neural Networks, San Diego.

27. Skrzypek, J., & Mesrobian, E. November 1987. Textual Segmentation: Gestalt Heuristics as a Connectionist Hierarchy of Feature Detectors. Paper presented at the Institute of Electrical and Electronics/Engineering in Medicine and Biology Society Annual Conference, Boston.

28. Skrzypek, J., Mesrobian, E., & Gungner, D. March 1988. Defining General Purpose Machine Vision: Metrics for Evaluation. Computer Science Department, UCLA.

29. Skrzypek, J., Mesrobian, E., & Gungner, D. April 1988. Machine Perception Laboratory Visual Task Sourcebook. Computer Science Department, UCLA.

Technology Assessment

30. Baker, E. L., Butler, F. A., & O'Neil, H. F., Jr. 1990. Proceedings of the Conference on Technology Assessment: Estimating the Future. Center for Technology Assessment, UCLA Center for the Study of Evaluation and Cognitive Science Laboratory, University of Southern California.

31. Baker, E. L., Butler, F. A., & O'Neil, H. F., Jr. 1991. Perspectives on Technology Assessment. Center for Technology Assessment, UCLA Center for the Study of Evaluation and Cognitive Science Laboratory, University of Southern California.

This is a collection of technical papers based on presentations made at the Conference on Technology Assessment: Estimating the Future. The list of papers follows.

#31 con.

Technical Papers

Models and Syntheses

Peled, Z., Peled, E., & Alexander, G. An Ecological Approach for Information Technology Intervention, Evaluation and Software Adoption Policies. Ben Gurion University, Israel.

Clark, R. E. Assessment of Distance Learning Technology. University of Southern California.

Kulik, J. Assessment of Computer-based Instruction. University of Michigan.

Assessment of Software Strategies

Moore, J. Assessment of Explanation Systems. University of Pittsburgh.

Swigger, K. M. Assessment of Software Engineering. University of North Texas.

Madni, A., & Freedy, A. Concurrent Engineering Technology Assessment. Perceptronics.

Examples of Training and Assessment Technologies

Lesgold, A. Assessment of Intelligent Training Technology. University of Pittsburgh.

Feurzeig, W. Tools for Scientific Visualization. BBN Laboratories.

Goldman, S., Pellegrino, J. W., & Bransford, J. Assessing Programs That Invite Thinking. Vanderbilt University.

Hawkins, J., Collins, A., & Frederiksen, J. Interactive Technologies and the Assessment of Learning. Bank Street College for Children and Technology.

Burns, H. Negotiated Topoi, Networked Epiphanies: Toward Future Technology Assessment Methods and Madness. University of Texas at Austin.

Braun, H. Assessing Technology in Assessment. Educational Testing Service.

Appendix

Artificial Intelligence Measurement System

1. Project Staff
2. Project Consultants

Artificial Intelligence Measurement System (AIMS)
Project Staff (1986-1990)

The following is the list of people who served as AIMS Project Staff at different times during the period of the contract. There was turnover from one academic year to another particularly with graduate students and support staff.

Project Management

Dr. Nancy Atwood -- Educational Psychology

Dr. Eva Baker -- Measurement; Learning and Instruction

Dr. Frances Butler -- Applied Linguistics

Dr. Dayle Hartnett -- Applied Linguistics; ESL instruction

Dr. Joan Herman -- Educational Evaluation; Measurement

Dr. Elaine Lindheim -- Educational Evaluation; Measurement

Project Support Staff

Kathleen Brennan -- Word Processor

Rory Constancio -- Office Manager

Elizabeth Freedman -- Secretarial Support

Katherine Frye -- Administrative Assistant

Wanetta Jones -- Conference Coordinator

Phyllis Kaelin -- Financial Affairs

Aeri Lee -- Administrative Support

Cindi Mercer -- Administrative Assistant

Sally Metry -- Administrative Assistant

Judy Miyoshi -- Administrative Assistant

Natural Language Understanding

Faculty and Staff

Dr. Eva Baker -- Measurement; Learning and Instruction

Dr. Frances Butler -- Applied Linguistics

Dr. Michael Dyer -- Artificial Intelligence; Natural Language Processing

Dr. Barbara Hecht -- Language Development

Dr. Walter Read -- Artificial Intelligence; Natural Language Processing

Dr. Merlin Wittrock -- Cognitive Psychology

Graduate Students

Tine Falk -- Learning and Instruction

Cheryl Fantuzzi -- Applied Linguistics

Richard Feifer -- Artificial Intelligence; Learning and Instruction

Susan Ferdman -- Computer Science; Learning and Instruction

Howard Herl -- Social Research Methods

Anat Jacoby -- Learning and Instruction

Younghee Jang -- Learning and Instruction

Karen Kellen -- Learning and Instruction

Emanuel Maidenberg -- Learning and Instruction

Patricia Mutch -- Linguistics

Mark Neder -- Applied Linguistics

Alex Quilici -- Artificial Intelligence

Regie Stites -- Linguistics; Anthropology

Eileen Terran -- Speech Pathology; Counseling Psychology

Jean Turner -- Applied Linguistics

Vision

Faculty and Staff

Dr. Josef Skrzypek -- Artificial Intelligence; Computer Vision

Graduate Students

Edmund Mesrobian -- Artificial Intelligence

David Gungner -- Artificial Intelligence

Paul Lin -- Artificial Intelligence

Emanuel Maidenberg -- Learning and Instruction

Eugene Paik -- Artificial Intelligence

Michael Stiber -- Artificial Intelligence

Expert Systems

Faculty and Staff

Dr. Eva Baker -- Measurement; Learning and Instruction

Dr. Harold F. O'Neil, Jr. -- Cognitive Science Laboratory, USC
(Subcontract)

Dr. Merlin Wittrock -- Cognitive Psychology

Graduate Students

Simon Chang -- Education

Anat Jacoby -- Learning and Instruction

Yujing Ni -- Learning and Instruction

John Novak -- Learning and Instruction

Dean Slawson -- Social Research Methods

Artificial Intelligence Measurement System

Project Consultants

Sourcebook (1988)

Jaime Carbonell, Computer Science Department, Carnegie Mellon University
Martha Evens, Computer Science Department, Illinois Institute of Technology
Evelyn Hatch, Applied Linguistics Department, UCLA
David Kieras, College of Engineering, University of Michigan
Carol Lord, Los Angeles IBM Scientific Center
Merlin Wittrock, Graduate School of Education, UCLA

Text Understanding (1990)

Carol Lord, Intelligent Text Processing, Inc., Santa Monica

Expert Systems (1987-90)

Ronald K. Hambleton, School of Education, University of Massachusetts
Zhongmin Li, School of Education, University of Southern California
Jason Millman, Cornell University
Harold F. O'Neil, Jr. (USC Subcontract)
Elliot Soloway, Department of Computer Science, Yale University
Kathleen Swigger, Computer Science Department, University of North Texas

Technology Assessment (1990)

Nancy K. Atwood, BDM International, Inc.
John D. Bransford, Vanderbilt University
Henry Braun, Educational Testing Service
Hugh Burns, University of Texas, Austin
Richard E. Clark, USC
William Doherty, BDM International, Inc.
Wallace Feurzeig, BBN Systems and Technologies Corporation
Susan F. Goldman, Vanderbilt University
Jan Hawkins, Bank Street College for Children and Technology
James Kulik, University of Michigan
Alan Lesgold, Learning R & D Center, University of Pittsburgh
Azad M. Madni, Perceptronics
Johanna Moore, University of Pittsburgh
Elad Peled, Ben Gurion University
Zimra Peled, Ben Gurion University
James W. Pellegrino, Vanderbilt University
Kathleen Swigger, Computer Science Department, University of North Texas

DISTRIBUTION LIST

Dr. Michael E. Atwood
NYNEX
AI Laboratory
500 Westchester Ave.
White Plains, NY 10604

Dr. Frances A. Butler
Center for the Study of Evaluation
145 Moore Hall, 405 Hilgard
University of California
Los Angeles, CA 90024

Dr. Richard Duran
Graduate school of Education
University of California
Santa Barbara, CA 93106

Dr. Patricia Baggett
School of Education
University of Michigan
610 E. University, Rm. 1302D
Ann Arbor, MI 48109-1259

CDR Robert Carter
Office of the Chief of Naval Oper.
OP-933D4
Washington, DC 20350-2000

ERIC Facility Acquisitions
2440 Research Blvd., Suite 550
Rockville, MD 20850-3238

Dr. Donald E. Bamber
Code 446
Naval Ocean Systems Center
San Diego, CA 92152-5000

Dr. Michelene Chi
Learning R&D Center
University of Pittsburgh
3939 O'Hara St.
Pittsburgh, PA 15260

LCDR Micheline Y. Eyraud
Code 602
Naval Air Development Center
Wesminster, PA 18974-5000

Dr. Isaac Bejar
Law School Admn. Services
P.O. Box 40
Newton, PA 18940-0040

Dr. Charles Clifton
Tobin Hall
Dept. of Psychology
Univ. of Massachusetts
Amherst, MA 01003

Dr. Marshall J. Farr, Consultant
Cognitive & Instructional Sciences
2520 No. Vernon St.
Arlington, VA 22207

Dr. Thomas G. Bever
Dept. of Psychology
University of Rochester
River Station
Rochester, NY 14627

Dr. Jere Confrey
Cornell University
Dept. of Education
Room 490 Roberts
Ithaca, NY 14853

Dr. Linda Flower
Carnegie-Mellon University
Department of English
Pittsburgh, PA 15213

Dr. C. Alan Bonetau
Department of Psychology
George Mason University
4400 University Drive
Fairfax, VA 22030

Dr. Lynn A. Cooper
Dept. of Psychology
Columbia University
New York, NY 10027

Dr. Alinda Friedman
Dept. of Psychology
University of Alberta
Edmonton, Alberta
Canada T6G 2E9

Sandra Borden
Naval Supply System Command
NAVSUP 5512
Washington, DC 20376-5000

Defense Technical Information Center
Cameron Station, Bldg. 5
Alexandria, VA 22314

Dr. Donald R. Gentner
Phillips Laboratories
345 Scarborough Rd.
Briarcliff Manor, NY 10510

Dr. Sam Glucksberg
Dept. of Psychology
Princeton University
Princeton, NJ 08540

Dr. William Howell
Chief Scientist
AFHRL/CA
Brooks AFB, TX 78235-5601

Dr. Demetrios Karla
GTE Labs, MS 61
40 Sylvan Rd.
Waltham, MA 02254

Dr. Susan R. Goldman
Peabody College, Box 45
Vanderbilt University
Nashville, TN 37203

Dr. Ed Hutchins
Intelligent Systems Group
Institute for Cognitive Science (C-015)
UCSD
La Jolla, CA 92093

Dr. J.A.S. Kelso
Center for Complex Systems
Building MT 9
Florida Atlantic University
Boca Raton, FL 33431

Dr. Timothy Goldsmith
Dept. of Psychology
University of New Mexico
Albuquerque, NM 87131

Dr. Janet Jackson
Rijkauniverstedt Grooingen
Biologisch Centrum, Vieugel D
Kerkiaan 30, 9751 NN Haren
The Netherlands

Dr. David Kieras
Tech. Communication Program
TIDAL, Bldg. 2360 Bonisteel Blvd
University of Michigan
Ann Arbor, MI 48109-2106

Dr. Sherrie Gott
AFHRL/MOMJ
Brooks AFB, TX 78235-5601

Dr. Robin Jeffries
Hewlett-Packard Labs, 3L
P.O. Box 10490
Palo Alto, CA 94305-0971

Dr. Alex Kirklik
Georgia Institute of Technology
Center for Human-Machine
Systems Research
Atlanta, GA 30332-0205

Prof. Edward Haertel
School of Education
Stanford University
Stanford, CA 94305

Dr. Peder Johnson
Dept. of Psychology
University of New Mexico
Albuquerque, NM 87131

Dr. Lois-Ann Kuntz
3010 SW 23rd Terrace, #105
Gainesville FL 32608

Dr. M. Holland
Army Research Institute for the
Behavioral and Social Sciences
5001 Eisenhower Avenue
Alexandria, VA 22333

Dr. Marcel Just
Carnegie-Mellon University
Dept. of Psychology
Pittsburgh, PA 15213

Dr. Jill F. Lehman
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Ms. Julie S. Hough
Cambridge University Press
40 W. 20th St.
New York, NY 10011

Dr. Michael Kaplan
Office of Basic Research
U.S. Army Research Institute
5001 Eisenhower Ave.
Alexandria, VA 22333-5600

Dr. Alan M. Lesgold
Learning R&D Center
University of Pittsburgh
Pittsburgh, PA 15260

Dr. Michael Levine
Educational Psychology
210 Education Bldg.
University of Illinois
Champaign, IL 61801

Dr. Barbara Means
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025

Dr. Donald A. Norman
C-015
Institute for Cognitive Science
University of California
La Jolla, CA 92093

Dr. Charlotte Linde
Structural Semantics
P.O. Box 707
Palo Alto, CA 94320

Dr. George A. Miller
Dept. of Psychology
Green Hall
Princeton Univ.
Princeton, NJ 08540

Library, NPRDC
Code P201L
San Diego, CA 92152-6800

Dr. Robert Lloyd
Dept. of Geography
Univ. of South Carolina
Columbia, SC 29208

Dr. Robert Mislevy
Educational Testing Service
Princeton, NJ 08541

Librarian
Naval Center for Applied Research
in Artificial Intelligence
Naval Research Lab., Code 5510
Washington, DC 20375-5000

Dr. Jane Malin
Mail Code EF5
NASA Johnson Space Center
Houston, TX 77068

Dr. Randy Munrow
Human Sciences
Westinghouse Science & Tech. Ctr.
1310 Beulah Rd.
Pittsburgh, PA 15235

Office of Naval Research
Code 114CS
800 N. Quincy Street
Arlington, VA 22217-5000

Dr. Elaine Marab
Naval Center for Applied Research
in Artificial Intelligence
Naval Research Lab., Code 5510
Washington, DC 20375-5000

Dept. of Administrative Sciences
Code 54
Naval Postgraduate School
Monterey, CA 93943-5026

Dr. Judith Orasanu
Basic Research Office
Army Research Institute
5001 Eisenhower Ave.
Alexandria, VA 22333

Dr. James L. McClelland
Dpet. of Psychology
Carnegie-Mellon University
Pittsburgh, PA 15213

Mr. J. Nelissen
Twente Univ. of Technology
Fac. Bibl. Toegepasla Onderwyskurde
P.O. Box 217, 7500 AE Enschede
The Netherlands

Dr. John Oriel
Navy Training Systems Center
Code 212
12350 Research Parkway
Orlando, FL 32826-3224

Dr. Kathleen McKeown
Columbia University
Dept. of Computer Science
450 Computer Science Bldg.
New York, NY 10027

Dr. A.F. Norcio
Code 5530
Naval Research Laboratory
Washington, DC 20375-5000

Dr. Glenn Osga
NOSC, Code 441
San Diego, CA 92152-6800

Dr. Ray S. Perez
ARI (PERI-II)
5001 Eisenhower Ave.
Alexandria, VA 22333

Nuria Sebastian
Dept. Palcologia Basica
Univ. Barcelona
Adolf Florence s.o.
08028 Barcelona, Spain

Dr. Ted Steinke
Dept. of Geography
Univ. of South Carolina
Columbia, SC 29206

Dr. Nancy N. Perry
Naval Education and Training
Program Support Activity, Code 047
Bldg. 2435
Pensacola, FL 32509-5000

Dr. Michael G. Shafto
NASA Ames Research Ctr.
Mail Stop 259-1
Moffett Field, CA 94035

Dr. Saul Sternberg
Univ. of Pennsylvania
Dept. of Psychology
3815 Walnut St.
Philadelphia, PA 19104-6196

Dr. Mary C. Potter
Dept. of Brain and Cognitive Sciences
MIT (E-10-039)
Cambridge, MA 02139

Dr. Valerie L. Shallin
Dept. of Industrial Engineering
State Univ. of New York
342 Lawrence D. Bell Hall
Buffalo, NY 14260

Dr. Thomas Sticht
Applied Behavioral
and Cognitive Science, Inc.
2062 Valley View Blvd.
El Cajon, CA 92019-2059

Dr. Stephen Reder
NWREL
101 SW Main, Suite 500
Portland, OR 97204

Dr. Ben Shneiderman
Dept. of Computer Science
University of Maryland
College Park, MD 20742

Mr. Michael J. Strait
UMUC Graduate School
College Park, MD 20742

Dr. Daniel Reisberg
Reed College
Dept. of Psychology
Portland, OR 97202

Dr. Randall Shumaker
Naval Research Laboratory
Code 5510
4555 Overlook Avenue, SW
Washington, DC 20375-5000

Dr. M. Martin Taylor
DCIEM, Box 2000
Downsview, Ontario
Canada M3M 3B9

Lt. Cdr Michael N. Rodgers
Canadian Forces Pers. App. Rsrch Unit
4900 Yonge St., Suite 600
Willowdale, Ontario M2N 6B7
Canada

Dr. Robert Smillie
Navy Personnel R&D
San Diego, CA 92132-6800

Dr. Zita E. Tyer
Dept. of Psychology
George Mason Univ.
4400 Univ. Drive
Fairfax, VA 22030

Dr. Fumiko Samejima
Dept. of Psychology
University of Tennessee
310B Austin Perry Bldg.
Knoxville, TN 37916-0900

Dr. James J. Staszewakj
Dept. of Psychology
Univ. of South Carolina
Columbia, SC 29210

Dr. Shih-sung Wen
Dept. of Psychology
Jackson State University
1400 J. R. Lynch St.
Jackson, MS 39217

Dr. Mark Wilson
School of Education
Univ. of California
Berkeley, CA 94720

Dr. Frank B. Withrow
U.S. Dept. of Education
Room 504D, Capitol Plaza
555 New Jersey Ave., NW
Washington, DC 20208

Dr. Wallace Wulfeck, III
Navy Personnel R&D Center
Code 51
San Diego, CA 92152-6800

Frank R. Yekovich
Dept. of Education
Catholic University
Washington, DC 20064

Dr. Joseph L. Young
National Science Foundation
Room 320
1800 G St., NW
Washington, DC 20650

Dr. Uri Zernik
Box 8
General Electric Research & Dev. Ctr.
Artificial Intelligence Program
Schenectady, NY 12301

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION The Regents of the University of California 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION Cognitive Science Program Office of Naval Research (Code 1142PT)
6c. ADDRESS (City, State, and ZIP Code) University of California, Los Angeles Office of Contracts and Grants Administration Los Angeles, California 90024	7b. ADDRESS (City, State, and ZIP Code) 800 North Quincy Street Arlington, VA 22217-5000
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Defense Advanced Research Projects Agency 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-86-K-0395
8c. ADDRESS (City, State, and ZIP Code) 1400 Wilson Boulevard Arlington, VA 22209-2308	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 61153N; PROJECT NO. RR04206; TASK NO. RR04206-0C; WORK UNIT ACCESSION NO. 442c022
11. TITLE (Include Security Classification) Artificial Intelligence Measurement System, Overview and Lessons Learned	
12. PERSONAL AUTHOR(S) Baker, Eva L. and Butler, Frances A.	
13a. TYPE OF REPORT Final	13b. TIME COVERED FROM 7/1/86 TO 2/28/91
14. DATE OF REPORT (Year, Month, Day) February 1991	15. PAGE COUNT 30
16. SUPPLEMENTARY NOTATION	

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP
12	05	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Artificial intelligence, natural language understanding, expert systems, machine vision, technology assessment

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This final report summarizes the work conducted for the Artificial Intelligence Measurement System (AIMS) Project which was undertaken as an exploration of methodology to consider how the effects of artificial intelligence systems could be compared to human performance. The research covered four primary areas of inquiry--natural language understanding, expert systems, machine vision, technology assessment. The four areas are discussed in turn with information provided regarding the goals of individual research efforts within each area. A list of project reports is included in the document.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. Susan Chipman	22b. TELEPHONE (Include Area Code) (703) 696-4318 22c. OFFICE SYMBOL ONR 1142CS

DD Form 1473, JUN 86 *Previous editions are obsolete.* SECURITY CLASSIFICATION OF THIS PAGE

S/N 0102-LF-014-6603